Welfare financing

Toolsema-Veldman, Linda; Allers, M.A.

university of groningen

faculty of economics and business

12004-EEF

Welfare financing: Grant allocation and efficiency

Linda A. Toolsema
Maarten A. Allers

SOM is the the research institute of the Faculty of Economics & Business at the University of Groningen. SOM has six programmes:
- Economics, Econometrics and Finance
- Global Economics & Management
- Human Resource Management & Organizational Behaviour
- Innovation & Organization
- Marketing
- Operations Management & Operations Research

Research Institute SOM
Faculty of Economics & Business
University of Groningen

Visiting address:
Nettelbosje 2
9747 AE Groningen
The Netherlands

Postal address:
P.O. Box 800
9700 AV Groningen
The Netherlands

T +31 50 363 7068/3815

www.rug.nl/som

Welfare financing: Grant allocation and efficiency

Linda A. Toolsema
University of Groningen
l.a.toolsema@rug.nl

Maarten A. Allers
University of Groningen

Welfare financing:
Grant allocation and efficiency

Linda A. Toolsema
University of Groningen, The Netherlands

Maarten A. Allers
COELO and University of Groningen, The Netherlands

July 2012

Abstract

Welfare is often administered locally, but financed through grants from the central government. This raises the question how the central government can prevent local governments from spending more than necessary. Block grants are more efficient than matching grants, because the latter reduce the local governments' incentive to limit welfare spending. However, conventional block grant financing is less equitable, indeed, it may put a heavy burden on local governments in economically weak regions. This paper considers block grants which depend on exogenous spending need determinants, and are estimated from previous period welfare spending. This allocation method gives rise to perverse incentives by reducing the marginal costs of welfare spending. We derive the conditions for such a grant to be more efficient than a matching grant, and apply our results to the Netherlands, where such a grant exists since 2004. We conclude that the Dutch style grant is likely to be more efficient than a matching grant. As it is also more equitable, other countries might want to consider introducing a similar grant.

Keywords: Welfare financing, Grant allocation, Block grant, Efficiency, Equity.

Corresponding author: L.A. Toolsema, Faculty of Economics and Business, University of Groningen, P.O.Box 800, 9700 AV Groningen, The Netherlands. E-mail: l.a.toolsema@rug.nl.

1. Introduction

In many countries, the payment of welfare benefits to the needy is a responsibility of subnational governments. Decentralization allows public services to be tailored to local preferences (Oates, 1972). Decentralization may also be more efficient (Hayek, 1945), as knowledge of local circumstances is needed to successfully run a welfare program. However, decentralized finance of redistributive programs is likely to break down as a result of the migration patterns it brings about ('race to the bottom'). Therefore, income redistribution is generally considered the responsibility of the central government. As a result, welfare is usually administered locally, but financed centrally. This raises the question as to how the center can induce local administrators to administer welfare efficiently in such a situation. In this paper, we interpret efficient administration as implementing programs to assist recipients in moving from welfare to work and carrying out fraud investigations in such a way that the number of welfare recipients is minimized.[1] With efficient administration, only those who really need it receive a welfare benefit. Thus, the policy question is: if the money for benefits is coming from elsewhere, what is to stop local administrators from being overly generous?

In the last decades, many countries have introduced some kind of welfare reform. Two important types of reform concern a change in financing (notably a shift from matching grants to block grants) and a decentralization of welfare policy (notably, of discretion over eligibility and welfare levels). The 1996 welfare reform in the US, which has attracted a lot of attention in the literature, combines both types of reform. Within European countries, however, regional differences in welfare eligibility and benefit levels are much less – or not at all – tolerated. In this paper we focus on the case of a *uniform* welfare policy, where local government behavior can be controlled only by grant allocation. The question then is: how to optimize the design of the grant to the decentralized governments which administer welfare?

[1] We ignore technical inefficiency by assuming that active labor market policies and fraud investigations are produced at minimum cost.

One way to provide incentives for efficiency is to provide local administrators with block grants rather than matching grants. However, the choice between matching grants and block grants involves a trade-off between efficiency and equity. Matching grants are not efficient because they reduce the costs to the local government of an extra welfare beneficiary. Therefore, they reduce the incentive of the local administration to keep welfare dependency at a minimum. On the other hand, matching grants are equitable because they guarantee that the central government shoulders an equal share of every local government's welfare burden. As a result, jurisdictions with high welfare spending needs due to exogenous circumstances receive a larger grant. In contrast, block grants are efficient, as they do not lower the cost of additional welfare recipients. But this comes at a price, as there is generally no guarantee that the welfare burden of every local government is shared by the central government to the same extent. Block grant financing may force local governments in economically backward regions to spend considerable sums of money from their own resources on welfare, while jurisdictions in affluent regions may not need to spend all their grant money on welfare.

Welfare is an entitlement program; people who qualify cannot be denied welfare. The grant system should reflect this. The challenge is thus to develop a grant allocation method which is both (sufficiently) efficient and (sufficiently) equitable. The Dutch 2004 welfare reform attempts to do so. An important feature of the Dutch system is that it aims at allocating block grants in such a way that municipalities which operate efficiently will not need to use own resources to finance welfare expenditures. At the same time, total grants add up to no more than forecasted aggregated welfare expenditures. If successfully applied, this should enable the Dutch to enjoy the benefits of block grants, without the disadvantage associated with them.

This paper analyzes the effects of welfare reform along the Dutch lines. Thus, we depart from the extensive literature on the optimal design of intergovernmental grants, but focus instead on the pros and cons of the specific type of grant system that is used in the Netherlands. The Dutch reform may be seen as an attempt to answer the question how local governments can be induced to administer welfare efficiently, while at the same time keeping eligibility and welfare levels uniform. The answer to

this question may also be relevant for other programs besides welfare, e.g., health or education programs.

The Dutch system seems to provide an attractive solution to promote efficient administration. We show, however, that the system has a weakness. In our basic framework, a local government decides on its inefficiency level, maximizing its objective function. In doing so, the local administrator balances marginal costs and benefits of welfare spending. As explained above, a matching grant is inefficient because it reduces the costs of an extra welfare beneficiary. That is, a matching grant directly decreases marginal costs of welfare spending – and thus of working inefficiently – and thereby affects the efficiency choice of local administrators, leading to a higher level of inefficiency. A standard block grant does not have this problem, and neither does the Dutch grant. However, the Dutch system makes *future* grants depend on current expenditures and thereby effectively reduces the marginal costs of spending too. Thus, like a matching grant, it affects the local governments' inefficiency decision by influencing the balance between marginal costs and benefits. The question then is how the Dutch style block grant compares to a matching grant in terms of efficiency.

It is important to realize that the Dutch situation is quite different from that in, e.g., the US. Eligibility rules are centrally determined, giving local administrators very little leeway in determining whether someone is entitled to receive benefits. The same is true for benefit levels. Even if there is some leeway, and therefore some minor differences across municipalities, mobility is limited in the Netherlands, in particular for people on welfare. Low income households depend on subsidized housing, for which there are considerable waiting lists. Thus, poor people do not migrate to obtain better welfare benefits, and a race to the bottom does not occur. Instead, the country faced the opposite problem that *too much* money was spent on welfare. This was the very reason for introducing the new block grant system, giving municipalities greater financial responsibility for their welfare expenditures and providing incentives for them to both decrease inflow and increase outflow of welfare (Van Es and Van Vuuren, 2010; Kok et al., 2007).

The remainder of this paper is organized as follows. Section 2 discusses related literature in several fields. Section 3 presents more detailed information about the Dutch welfare grant allocation method. Section 4 describes and solves a model of the efficiency choice at the municipality level. Section 5 adds to this several types of grant allocation and studies their effects on efficiency. Section 6 concludes.

2. Related literature

This paper is related to several strands of literature: those concerning welfare reform, intergovernmental grants, fiscal equalization, mechanism design, and yardstick competition.

Analyses of *welfare reforms* often simultaneously deal with both a change in financing and a decentralization of discretion over welfare policy.[2] Moreover, empirical studies of the influence of financing arrangements typically do not discriminate between the effects on benefit levels, which do not concern us here, and on the number of recipients. An exception is Baicker (2005), who uses US data from 1948 to 1963 to separately estimate the effect of the match rate of the federal grant on welfare benefits per recipient and on the number of welfare recipients. For the former, she found a price elasticity of around -0.4; for the latter, an elasticity of around -0.3. Thus, a matching grant results in a higher number of welfare beneficiaries than a block grant, which has a match rate of zero.

That is in line with the traditional theory of *intergovernmental grants*, where the differential effects of matching grants and block or lump-sum grants have been discussed extensively (for a review, see, e.g., Wildasin, 1986; Oates, 1972, 1999; or Bird and Smart, 2002). The upshot of this theory is that a matching grant, by lowering the marginal cost of public services, has a greater stimulating effect on local spending than does a lump-sum grant of the same amount. Matching grants may be optimal if local decision making produces inefficient outcomes, e.g., in the case of externalities.

[2] See, e.g., Chernick (1998), Ribar and Wilhelm (1999) and Blank (2002) for the US welfare reform of 1996, and Gilbert and Rocaboy (1996) for the 1994 reforms in France.

If that is not the case, unconditional block grants are the most efficient grants, as they do not distort local governments' spending decisions.

More recent studies of intergovernmental grants stress that this conclusion only holds under conditions of full information and unlimited capacity on the part of the central government to commit itself to grant policy. If local governments expect that the central government will bail them out ex post with extra grants, a moral hazard problem occurs, and local governments are likely to overspend (e.g., Goodspeed, 2002). This soft budget constraint literature is closely related to the literature on decentralized leadership (e.g., Köthenbürger, 2004; Akai and Sato, 2008; Breuillé et al., 2010). In these studies, local jurisdictions make their taxing and spending decisions ex ante, and the central government decides on grant allocation ex post.

In our model, however, there is no soft budget constraint, and the grant allocation system is determined ex ante. We do have asymmetric information, though. If local governments are to give assistance to anyone who is eligible according to centrally determined rules, and if benefit levels are determined centrally as well, local governments need sufficient revenues to pay out benefits. Exogenous determinants of welfare dependency (e.g., health, education, labor market) differ considerable between jurisdictions. The welfare block grant allocation should account for this. This touches upon the literature on *fiscal equalization*. In many countries, fiscal disparities are equalized to some extent through a system of intergovernmental grants.[3] Equalization of spending needs requires quantifying them, which is notoriously difficult (Duncan and Smith, 1996). One of the techniques that may be employed is a regression of spending on cost determinants (Ladd, 1994; Bradbury and Zhao, 2009). This technique is used in the Netherlands to derive the welfare grant allocation formula.

[3] Equalization has been advocated on the grounds that it improves locational efficiency (Buchanan, 1950, 1952; Buchanan and Goetz, 1972; Boadway and Flatters, 1982); on equity grounds (Le Grand, 1975; Bramley, 1990); as an insurance against regional shocks (Bucovetsky, 1998; Von Hagen, 2006) and in order to improve transparency and thereby facilitate the local decision making process (Allers, 2012). For a review of the arguments for equalization, see Boadway (2006).

However, asymmetric information limits the central government's ability to design an optimal grant ex ante (e.g., Raff and Wilson, 1997; Cornes and Silva, 2002; Huber and Runkel, 2006; Breuillé and Gary-Bobo, 2007). Like these studies, we analyze a model where the central government cannot directly observe whether a local government has high or low costs. However, in our case, cost disparities among local governments can be estimated. This estimate is biased because local government efficiency levels are unobserved and thus omitted from the regression. As the grant allocation system provides an incentive to reduce inefficiency, this bias may decrease over time. Thus, our paper is related to the *mechanism design* literature, where, e.g., Riordan and Sappington (1988) and Strausz (2006) show that the informational rent earned by an agent can be reduced if more accurate information about that agent becomes available ex post. Akai and Silva (2009) also analyze a model where ex post information enables the center to improve the grant system, but, in their model, ex post information about cost levels is complete, whereas, in our model, it is biased. Moreover, their model includes soft budget constraints.

A final related strand of literature concerns *yardstick competition*, an instrument to provide incentives for regulated monopolies (e.g., Shleifer, 1985). This instrument may be applied if the regulator does not know the minimal cost at which a firm can produce. Yardstick competition entails comparing similar regulated firms with each other. For any given firm, the regulator uses the costs of comparable firms to infer the attainable cost level. The regulator may use this information to let each firm compete with its own shadow firm. Thus, each firm has an incentive to lower costs as much as it can. Note that this requires that the regulator observes actual cost levels of firms. In our case, however, inefficiency is unobserved.[4]

[4] Another difference is that, in Shleifer's model, firms maximize profits minus adjustment costs in a one-shot game, whereas in our model the maximization problem is more complicated. Also, in our model, the transfer or grant is more complicated than the equilibrium transfers derived by Shleifer.

3. Welfare finance in the Netherlands

The territory of the Netherlands is divided into 441 (in 2009) local governments, or municipalities. Municipalities are responsible, among other things, for administering welfare. Eligibility rules and welfare benefit levels are uniform across the country. Until 2001, each municipality financed 10 percent of its welfare benefits from its own coffers, while 90 percent was reimbursed by the central government through an open-ended matching grant. Clearly, this did not provide a strong incentive to limit welfare payments by helping recipients find work or by clamping down on fraud. In order to improve this incentive, the match rate was reduced from 90 to 75 percent in 2001. As from 2004, no reimbursement takes place any more. Matching grants have been replaced by block grants. If a municipality spends more than its block grant, it bears the extra expenditures itself (up to a point, see below). If it spends less, it may use the balance as it sees fit. Figure 1 summarizes the reform. The line AB indicates the hypothetical budget constraint without grants. By lowering the price of an additional beneficiary, the open-ended matching grant used before 2004 rotated the local budget constraint from AB to AC. The block grant used since 2004 shifts the budget constraint from AB to DE. DE is steeper than AC, reflecting the higher price of an additional welfare recipient.

[INSERT FIGURE 1 ABOUT HERE]

In addition to benefit payments, municipalities also incur administrative costs. We define administrative costs as the costs of running a welfare program over and above the welfare benefit payments themselves. Administrative costs include, inter alia, the costs of establishing eligibility, of helping welfare recipients find a job (e.g., work programs), and of fraud investigations. Administrative costs are paid partially out of an earmarked block grant, and partially out of own resources.[5] Own resources include

[5] Every municipality receives a block grant ("participatiebudget") earmarked for helping unemployed persons find work, for integrating immigrants and for educating adults with insufficient schooling. The grant a municipality receives depends on the number of welfare recipients, the number of unemployment benefits, the size of the work force and an indicator for the availability of low-skilled jobs (*Besluit participatiebudget*, Annex 1; available – in Dutch – through www.overheid.nl). Although

a (considerable) equalizing unconditional lump-sum grant from the central government and (comparatively modest) local tax revenues.

The new financing arrangement introduced in 2004 was accompanied by greater local autonomy in the treatment of welfare recipients. However, it is important to stress that this new autonomy is limited to administration. Local governments have discretion over the programs they employ to assist recipients in moving from welfare to work, and the intensity of their fraud investigations. Eligibility rules and welfare benefit levels are still uniform and determined centrally. This is an important difference with the 1996 welfare reform in the US. Municipal welfare expenditures[6] can only be lowered by reducing caseloads. As assistance to the truly needy cannot be refused, caseloads can only be reduced by weeding out fraudulent beneficiaries and by helping *bona fide* recipients find work.

The nationwide budget available for welfare block grants, referred to as the *macro budget*, is calculated annually based on forecasts of the number of persons eligible for welfare. These forecasts are made by the independent Netherlands Bureau for Economic Policy Analysis, known by its Dutch acronym CPB.[7] Forecasts are based on the number of welfare beneficiaries, the development of the number of unemployed in the previous years,[8] and regulatory changes that may affect welfare volumes.

The macro budget is allocated over municipalities according to the following rules. For small municipalities (fewer than 25,000 inhabitants, where 9 percent of welfare

---

this grant increases when the number of welfare recipients goes up, the money is earmarked. Unlike the grant aimed at financing local welfare benefits, it cannot be used for other purposes. Therefore, we assume it does not enter the local government's utility function, and we ignore this grant in the following sections.

[6] In this paper, 'welfare expenditures' refers to welfare benefit payments only. That is, welfare expenditures do not include administrative costs.

[7] See www.cpb.nl.

[8] People losing their job normally are entitled to an unemployment benefit for a period which depends on their employment history. After this period, they may apply for a (usually lower) welfare benefit if they have insufficient means to support themselves and their families.

recipients live), the share of the macro budget in year *t* depends on their share of welfare expenditures in year *t*-2. For large municipalities (40,000 inhabitants and more, 80 percent of welfare recipients), a formula applies, which includes both demographic and labor market characteristics. The allocation formula is updated regularly. Because a formula that covers smaller municipalities reasonably well could not be derived, this method does not apply to them. For medium-sized municipalities, a hybrid system applies: their share is partly derived from their expenditure share in year *t*-2, and partly from the formula.

It has proven difficult to derive a stable allocation formula. Municipalities may see their calculated share of the macro budget rise or fall considerably from one year to the next. In order to insulate local governments from budgetary shock too great to cope with, differences between the block grant and actual welfare expenditures are limited both *ex post* and *ex ante*. These limits are analyzed in Appendix B.

Presently, the welfare grant allocation formula contains 14 variables.[9] Among these are the number of single parent households, the number of lowly educated people, employment growth in the region to which the municipality belongs, and the number of disability benefits. The weights of these variables are derived annually[10] from a regression at the municipal level of welfare expenditures on the determinants included in the formula.

This approach is not without problems. That is because municipalities operate at different levels of efficiency. Actual welfare expenditures are a biased indicator of spending need, which is defined as the welfare spending a municipality would incur if it operated efficiently (as defined above). Greater efficiency in the past results in lower welfare expenditures, which translates into lower weights in the formula for the variables on which the municipality scores relatively high, and therefore into a lower grant. As a result, bad behavior in the past is rewarded. This provides perverse

[9] We describe the Dutch system as it existed in 2009. The grant allocation formula may be found in Annex 1 of *Besluit WWB 2007*, which is available (in Dutch) through www.overheid.nl.

[10] In practice, the grant formula is left unchanged in some years.

incentives and distorts efficiency. However, the allocation formula is updated regularly, and policymakers expect that, as the new grant design improves efficiency across the board, this bias will gradually disappear. In the next two sections we will show that this is not to be expected. We do so by presenting a theoretical framework which describes the effects of different grant allocation methods on the local governments' efficiency decision.

4. Local governments' efficiency decision

In this section we focus on the choice of the efficiency level by the local authorities. As already explained, we interpret efficiency as implementing active labor market policies and fraud investigations in such a way that benefit payments are minimized: benefits are only given to those who really need it.[11] For now, we use a very general function to describe the grant allocation method. We will turn to specific allocation methods in the next section.

4.1. Model

We assume that the efficiency decision is not only based on a local government's expenditures on welfare and on grant allocation, but also on some 'easy life function', which describes the monetary equivalent of the utility that the local government's administrators derive from working inefficiently. This utility may, e.g., take the form of political gains that may be derived from handing out benefits generously, or it may simply reflect the utility of leaning back and not exerting too much effort on work programs or enforcement. Furthermore, we assume that the local government takes into account how actual welfare expenditures will depend on the level of inefficiency. That is, although the central government does not observe the local governments' inefficiency levels, the local government has full information. Finally, we impose a maximum inefficiency level,[12] which should be interpreted as follows. Although the central government cannot observe the inefficiency level, it will notice when a local

[11] Thus, we ignore technical inefficiency. Many municipalities contract out programs to help welfare recipients find work to private firms which operate in more than one municipality. Therefore, this assumption does not seem to be unduly unrealistic there.

[12] The assumption of a maximum inefficiency level does not qualitatively affect the results. It merely avoids the possibility of extreme inefficiency which does not seem to make sense in practice.

government 'misbehaves' in an extreme way, and it will intervene and impose a severe punishment.

As a benchmark we first consider the case of an open-ended matching grant. Suppose that the central government reimburses a share 1-α of a local government's welfare payments, leaving only a share α to be paid out of the local government's own resources. In addition, we assume that the local government also pays the administrative costs out of its own resources. This yields the following maximization problem for the local government under the matching grant, which is indicated by a subscript 0 denoting the benchmark situation:[13]

$$
\begin{aligned}
P_0: \ & \max_{Z_0} -\alpha Y_0 - C(Z_0 \mid X) + L(Z_0) \\
\text{(1)} \qquad & \text{s.t.} \ \ Y_0 = X\beta + Z_0\gamma \\
& \qquad Z_0 \leq Z^{\max}.
\end{aligned}
$$

Here, Z_0 denotes the inefficiency level of the local government, $Z_0 \geq 0$ and greater Z_0 means greater inefficiency. The lower Z, the greater effort is exerted by the administrators to reduce welfare expenditures. Y_0 denotes the local government's welfare expenditures and is determined both by Z_0 and by the exogenous spending need determinants X (a 1xn vector).[14] We let X be time-independent for expositional convenience. In reality these variables may change over time, but they do so only gradually, and they cannot be influenced by the welfare administrators. $C(Z_0|X)$ is the administrative cost function[15] with $C(Z|X) \geq 0$, $C'(Z|X) < 0$, and $C''(Z|X) \geq 0$. $L(Z_0)$ is the easy life function with $L'(Z) > 0$, and $L''(Z) \leq 0$; $Z^{\max}$ is the maximum inefficiency level. Finally, α, β (an nx1 vector), and γ are parameters with $\gamma > 0$. Note that 1-α is the match rate of the welfare grant to the local government, with $\alpha \in [0,1]$.

[13] In this section, we focus on a single local government and therefore do not use an index to denote the identity of the authority.

[14] In reality, past inefficiency levels may also play a role. When the administrator helps a person to find a job, he or she may still have that job next year, so the administrator continues to enjoy lower expenditures. We ignore this in the theoretical framework.

[15] We use the notation $C(Z_0|X_0)$ rather than $C(Z_0,X_0)$ in order to simplify the notation of the derivative of C with respect to Z (writing C'), taking into account that X is an exogenous variable.

In the analysis below, only the difference between the functions C and L (or between their derivatives) matters. Thus, mathematically, they play the same role and one could argue that one of the two is redundant. However, from an economics point of view, the two functions have distinct interpretations. Therefore, we choose to explicitly use both C and L below.

The assumptions regarding the administrative cost function and the easy life function can be interpreted as follows: more inefficiency (higher Z) makes administrators' lives easier, but it does so at a decreasing rate. More inefficiency also lowers administrative costs, e.g., because of less effort to help beneficiaries find work, but it does so at a decreasing rate. Note that we focus on a single period in this maximization problem. Including future periods in the objective function (as we do below) would not affect the solution for the problem under the matching grant, P_0, however, and therefore we ignore those for expositional convenience.

It is important to note that the central government can observe welfare expenditures Y and spending need determinants X, but not the inefficiency level Z. Also, the parameter β is not observed by the central government. Although X is assumed to be constant, the parameter β may change over time as macro-economic conditions vary. The central government cannot infer Z from the observables.

Now consider a block grant system along the lines of the Dutch welfare reform. We use a time subscript $t \geq 1$ because future periods do matter under this system. The block grant for year t depends on last year's welfare expenditures of all local governments together and on macroeconomic variables, which together determine the macro budget, as well as on a grant allocation formula. Consequently, a local government can only influence the grant in year t via its inefficiency level in the previous year, Z_{t-1}. The local government has to pay the full welfare expenditures out of the grant, supplemented with the local government's own resources if necessary. Thus, there is no reimbursement anymore. Administrative costs and the easy life hypothesis remain as before. Thus, under the block grant system at time $t \geq 1$, the local government solves the following problem:

$$P_t: \max_{Z_t} \sum_{\tau=t}^{\infty} \delta^{\tau-t}\left[B_\tau(Z_{\tau-1}) - Y_\tau - C(Z_\tau \mid X) + L(Z_\tau)\right]$$

(2) $$\text{s.t. } Y_\tau = X\beta + Z_\tau\gamma$$
$$Z_\tau \le Z^{\max} \quad \forall \tau = t, t+1, \ldots$$

Here, we define $\delta \in [0,1]$ to be the discount factor. B_τ represents the grant at time τ, where we assume $B_\tau'(Z_{\tau-1}) \ge 0$. This assumption implies that inefficiency is rewarded by a larger grant in the next period. We also assume that the greater the local government, the larger the effect of its behavior on its future grant. Note that formally, B_τ is a function of $Y_{\tau-1}$, which itself is a function of $Z_{\tau-1}$. We simplify this by writing B_τ as a function of $Z_{\tau-1}$. Also for simplicity, we assume that local governments differ only in their *Z*, *X*, and *Y* and *B*; not in their functions *L* and *C* or parameters α, β, γ, and δ.

In this section, we thus use a general grant function B_τ, although we put some restrictions on it. With this very general allocation rule, we can already derive our main results. In Section 5 we will study the effects of specific allocation methods in more detail. These methods do satisfy (and indeed inspired) the restrictions on the function B_τ in this section. For example, with a macro budget which is determined in part by previous welfare expenditures (see Section 5), higher inefficiency ceteris paribus implies a larger grant in the next period as a result of a larger macro budget (and, depending on the grant allocation system, possibly a larger share of the macro budget as well). Also, in such a setting, the extreme case of a very small (in terms of welfare expenditures) local government may have $B_{\tau+1}'(Z_\tau) \approx 0$, as its behavior may not affect its future grant, because it is too small to affect the macro budget or the budget shares.

4.2. Solution

In the benchmark case, under a matching grant, the local government solves the problem P_0 in (1). This yields the first-order condition (FOC)

(3) $$L'(Z_0) - C'(Z_0 \mid X) = \alpha\gamma.$$

Depending on the shapes of the functions L and C, this may of course yield a corner solution at either $Z_0 = 0$ or $Z_0 = Z^{\max}$. We assume in the following that the FOC (3) has an interior solution denoted by Z_0^*. Note that Z_0^* depends on X via the administrative cost function. This implies that different local governments – that is, with different X – will choose different efficiency levels even if the functions C and L are the same across local governments.

The solution is shown graphically in Figure 2. Figure 2 plots the difference between the derivatives of the easy-life function L and the administrative cost function C, assuming for simplicity that this difference L'-C' is linear. Given our assumptions, L'-C' is positive and decreasing in Z. The Figure shows how to derive the equilibrium values of the inefficiency level Z. Under the benchmark matching grant, equilibrium occurs at the point where L'-C' equals $\alpha\gamma$, a constant. The corresponding equilibrium is illustrated by E_0 in the Figure. In the extreme case of a match rate of one, $\alpha\gamma = 0$, and equilibrium occurs at $Z^{\max}$. With a fixed block grant, or without any grant, the match rate is zero, so $\alpha\gamma = \gamma$, and the local government chooses efficiency level Z^{**} (E_2 in Figure 1). Note that Z^{**} is the lowest efficiency level that can be reached by changing the grant system. This equilibrium arises if local governments have no influence whatsoever over the grant they receive. For expositional convenience, and because perfect efficiency seems unlikely even in this case, we assume that $Z^{**} > 0$.

[INSERT FIGURE 2 ABOUT HERE]

Under the block grant the local government solves the problem P_t in (2). The corresponding FOC is

$$L'(Z_t) - C'(Z_t \mid X) = \gamma - \delta B_{t+1}'(Z_t). \tag{4}$$

This differs from the FOC of the benchmark model, (3), because now the match rate is zero (α=1) and there is a block grant B_{t+1} which depends on Z_t. Again, we assume an interior solution, and again the equilibrium efficiency level depends on the exogenous variables in the allocation formula, so different local governments (with

different X) will choose different efficiency levels even if the functions C and L are the same across local governments.

4.3. Results

Now consider what happens if a matching grant is replaced by a block grant. With a matching grant, the FOC (3) of the benchmark model shows how the local government balances marginal benefits and costs of inefficiency, i.e., of welfare spending. Marginal benefits consist of increased easy life and reduced administrative costs (left-hand side of FOC), and marginal costs are reflected by the increase in net-of-grant welfare expenditures (right-hand side or RHS of FOC). Under the block grant, the FOC is given by (4). Again, the RHS can be interpreted as the marginal cost of inefficiency. It describes the effect of Z_t on the local government's welfare expenditures, γ, minus the present value of the budget increase in the next period. Together, this can again be interpreted as the effect on net-of-grant welfare expenditures.

Comparing (3) and (4) reveals the following. First, a matching grant (α<1) reduces the costs of an extra welfare beneficiary from γ to $\alpha\gamma$. That is, a matching grant decreases marginal costs of welfare spending – and thus of working inefficiently – and thereby affects the efficiency choice of local administrators, leading to a higher level of inefficiency. Second, the block grant as described above makes future grants depend on current inefficiency (via expenditures) and thereby reduces the marginal costs of spending as well, this time to $\gamma - \delta B_{t+1}'(Z_t)$. Essentially, in this case moral hazard arises because local governments have an incentive to reduce their efficiency in order to get a higher grant in the future. Thus, both the matching grant and the block grant with dependence on previous inefficiency affect the local governments' inefficiency decision in a similar way: by influencing the balance between marginal costs and benefits of inefficiency.

Figure 2 illustrates how the block grant equilibrium Z_t^* compares to the benchmark matching grant equilibrium Z_0^*. At time $t \geq 1$, the FOC (4) not only involves the functions L and C, but also the function B_{t+1}, where $B_{t+1}'(Z_t) > 0$. Equilibrium is illustrated by E_1 in the Figure. Note that although in Figure 2 we have drawn

$\gamma - \delta B_{t+1}'(Z_t)$ as a decreasing function of Z_t, it could alternatively be increasing (or even nonmonotonic) depending on the sign of $B_{t+1}''(Z_t)$.[16] Using Figure 2 we obtain the following result.

Proposition 1: Moving from a matching grant to a block grant may induce a local government to increase efficiency ($Z_t^* < Z_0^*$), depending on parameter values.

Note however that, depending on the shapes of L', C', and B_{t+1}', and the values of the parameters α, γ, and δ, the equilibrium inefficiency level may in fact increase rather than decrease with the introduction of a block grant system. The FOC (4) associated with the new system can be rewritten as

$$L'(Z_t) - C'(Z_t \mid X) - \gamma + \delta B_{t+1}'(Z_t) = 0.$$

We can evaluate the left-hand side of this FOC in the benchmark equilibrium inefficiency level Z_0^* (i.e., substituting the FOC (3) for time t=0) as

$$\alpha\gamma - \left\{\gamma - \delta B_{t+1}'\left(Z_0^*\right)\right\}.$$

As can also be seen from Figure 2, the inefficiency level Z will decrease relative to Z_0^* with the introduction of the new system if this expression is negative ($\gamma - \delta B_{t+1}'\left(Z_0^*\right) > \alpha\gamma$), but it will increase instead if the expression is positive. The expression is increasing in α and δ and decreasing in γ. Thus, for the block grant system to indeed enhance efficiency, we need both the match rate under the old system (1-α) and the effect of a local government's inefficiency on its welfare expenditures (γ) to be sufficiently large, and the discount factor (δ) to be sufficiently small. Also, since $B_{t+1}' \geq 0$, the function B_{t+1} should not be too steep.

[16] Note that if X does depend on t then the curve L'-C' would shift when we move from the matching grant equilibrium (E_0) to the block grant equilibrium ($t \geq 1$; E_1 or E_2) since in general C' depends on X.

For smaller local jurisdictions, the effect of increased efficiency on the macro budget is smaller than for large jurisdictions. Therefore, increased efficiency will reduce their next period grant by a smaller amount (they face a less steep B_{t+1} function). Therefore inefficiency levels are more likely to decrease for relatively small local governments (see also Section 5.1), and in general in settings with many local governments – since that implies that individual local governments will be smaller relative to the total welfare expenditures, at least on average. In the case of an extremely small local government, $B_{t+1}'(Z_t) \approx 0$, and we end up in equilibrium E_2 in Figure 2.

Proposition 2: The efficiency boost from replacing a matching grant by a block grant will decrease with local government size.

Thus, in this model, if parameter values are such that the introduction of a block grant *decreases* efficiency for some local governments, this will be the case for relatively large jurisdictions. Section 5.1 illustrates this. Of course, this result crucially depends on our assumption that a larger local government's behavior has a stronger effect on its future grant.

From inspection of the FOCs and Figure 2 it is also easy to derive the following result.

Proposition 3: Under the block grant complete efficiency (Z=0) will not necessarily obtain, depending on parameter values.

The most efficient grant is a fixed block grant, equivalent to a matching grant with α = 1. With such a grant, equilibrium occurs at Z^{**}, which is still higher than zero except when $\gamma \geq L'(Z_t) - C'(Z_t \mid X) \; \forall \, Z_t$. So, in general, we have both less than perfect efficiency, and continuing disparities in efficiency across local governments.[17]

[17] Recall that size differences are not the only reason why inefficiency levels will differ across local governments. Efficiency levels are also determined by the exogenous variables X, via the administrative cost function C.

Due to the setup of our model, convergence to the new equilibrium after introduction of the new system is immediate. If the derivative B_{t+1}' does not depend on t, then the new equilibrium inefficiency level Z_t^* is in fact independent of t. However, if the derivative B_{t+1}' does depend on t (or, contrary to our assumption, we have time-dependent exogenous spending need determinants X_t), the equilibrium inefficiency level Z_t^* induced by the new system is also time-dependent. Such dependence of B_{t+1}' could for example be due to the fact that the macro budget and its response to an individual local government's inefficiency level depend also on other local governments' inefficiency levels. If this time-dependence of B_{t+1}' reduces over time, for example because inefficiency levels decrease, this could result in some kind of transition path towards a new equilibrium inefficiency level.

The model can be extended to include possible loss aversion. The municipality's objective function may put a greater weight on a deficit than on a surplus. We analyze this extension of the model in Appendix A.

5. Grant allocation

The function B_{t+1} is part of the design of the welfare allocation model, and therefore can be influenced by the policy maker. We now turn to a discussion of the implications of some specific allocation models.

We begin by considering a simple hypothetical system where every local government's share in the macro budget is constant. Then we analyze two simplified systems which are based on the arrangement that is in place in the Netherlands, for small and large municipalities, respectively. First, we consider grant shares which equal previous expenditure shares. Second, we analyze a model where grant shares are based on a regression of welfare expenditures on exogenous variables reflecting spending need. For simplicity, we set the macro budget equal to total welfare expenditures in the previous period. In this section we index local governments by a subscript i, i=1,...,m.

5.1. Fixed shares

Suppose that each local government receives a fixed share of the macro budget. By decreasing Z_i, a local government receives the benefits of greater efficiency while sharing the cost in terms of a reduced grant in the next period (resulting from a lower macro budget) with all other local governments.

The grant for local government i in period $t \geq 1$ is given by

$$(5) \qquad B_{i,t} = \theta_i \sum_{i=1}^{m} Y_{i,t-1},$$

where θ_i is the fixed share of local government i in the macro budget. The θ_i's are exogenous parameters and are assumed to be independent of t in this subsection, with $\theta_i \in [0,1], \sum_{i=1}^{m} \theta_i = 1$. For example, they could be determined as historical shares by $\theta_i = Y_{i,0} / \sum_{i=1}^{m} Y_{i,0}$. Note that if a local government increases its expenditures by one euro, its grant for next year increases by $\theta_i \leq 1$ euros. We now have $B_{i,t+1}'(Z_{i,t}) = \theta_i \gamma$, and the FOC (4) becomes

$$L'(Z_{i,t}) - C'(Z_{i,t} \mid X_i) = \gamma(1 - \delta\theta_i).$$

Recall that the right-hand side (RHS) of the FOC was equal to $\alpha\gamma$ with a matching grant (FOC (3)). Now it is again a constant and local government i increases efficiency after the introduction of the block grant system if and only if $1 - \delta\theta_i > \alpha$, or, equivalently, $\delta\theta_i < 1 - \alpha$. Here, $\delta\theta_i$ is the present value of the grant increase in the following year resulting from spending one additional euro on welfare under the fixed shares block grant system, while $1 - \alpha$ represents the grant increase resulting from spending one additional euro on welfare under the matching grant system.

Proposition 4: A block grant with fixed shares θ_i entails $B_{i,t+1}'(Z_{i,t}) = \theta_i \gamma$. This block grant is more efficient than a matching grant if and only if $\delta\theta_i < 1-\alpha$. This is more likely for local governments with a low share θ_i of the macro budget.

This clearly illustrates the result presented in Section 4 that, ceteris paribus, large local governments (those with greater θ_i, for example due to their large share in historical welfare expenditures) will have greater inefficiency under the block grant system. Large local governments therefore are more likely than small local governments to decide to increase rather than decrease their inefficiency level after the introduction of the block grant (Proposition 2).

5.2. Grant based on previous period share

Now suppose that the grant share depends on a local government's share in welfare expenditures in the previous period. Thus, $\theta_{i,t} = Y_{i,t-1} / \sum_{i=1}^{m} Y_{i,t-1}$. In this case, $B_{i,t}$ depends on $Z_{i,t-1}$ not only because $Z_{i,t-1}$ influences the macro budget, but also because it now influences the local government's share of the macro budget.

Substituting the expression for $\theta_{i,t}$ into the expression for the grant of local government i, (5), immediately yields $B_{i,t} = Y_{i,t-1}$. Given our assumption about the determination of the macro budget (i.e., the macro budget equals total welfare expenditures in the previous period), a local government's grant for year t simply equals its expenditures in the year before. Thus, each euro of expenditures directly translates into one euro grant for next year. This implies $B_{i,t+1}'(Z_{i,t}) = \gamma$. Note that with fixed shares (the previous subsection) this derivative is multiplied by the share θ_i, which will in general be small. Thus, this derivative is much larger with grants based on the previous period's share than it is with fixed shares. The right-hand side of the FOC (4) now equals $(1-\delta)\gamma$ and is much smaller than with fixed shares, implying that we now have far greater $Z_{i,t}$ in equilibrium.

Proposition 5: With previous period shares we have $B_{i,t+1}'(Z_{i,t}) = \gamma$ and the block grant system is more efficient than a matching grant if and only if $\delta < 1-\alpha$. This is much less likely than with fixed shares.

This result is easily understood. In the condition $\delta < 1-\alpha$, the δ represents the present value of the grant increase in the following year resulting from spending one additional euro on welfare under the fixed shares block grant system, while $1-\alpha$ represents the grant increase resulting from spending one additional euro on welfare under the matching grant system.

5.3. Grant based on regression

If block grants are used but equity is a concern, past expenditures are probably not the best instrument to improve equity. With exogenous spending need determinants observable to all parties, econometric techniques allow forecasting future spending needs and allocating the available budget accordingly. We now consider such a sophisticated method where grant shares are derived from a regression of welfare expenditures on exogenous spending need determinants. There is, however, one problem with this method. As reflected in the model from the previous subsection, there is an additional explanatory variable, inefficiency, which cannot be observed. In practice, this variable is ignored when forecasting spending need. Below we analyze how this omitted variable problem affects grant shares and efficiency.

In order to formalize this, we first consider the 'true model' relating Y to X,

$$Y_t = X\beta + Z_t\gamma + \mu_t, \tag{6}$$

using matrix notation. Here, Y_t, Z_t and μ_t are mx1 vectors, with m the number of local jurisdictions, X is an mxn matrix with n the number of exogenous spending need determinants, and β (nx1) and γ (scalar) are parameters. This is the same equation relating Y to X and Z as before, see (1), but now with an i.i.d. disturbance term added (we assume that $E\mu_t = 0$). For a truly fair grant allocation, one would need to know the parameter β. However, since Z_t is unobservable, β cannot be estimated. The regression model used is therefore an approximation:

$$Y_t = X\varphi + \varepsilon_t, \tag{7}$$

assuming that the disturbance term is i.i.d. with $E\varepsilon_t = 0$.[18] Clearly, the estimate $\hat{\varphi}$ which results from this estimation is a biased estimate of β, unless X and Z_t are orthogonal, which is highly unlikely. [19]

The model (7) is re-estimated every year, so the estimate of φ changes annually . This is indicated by a subscript t. The estimate for φ calculated at time t is given by

$$\hat{\varphi}_t = (X'X)^{-1}X'Y_t,$$

with $\hat{\varphi}_t$ an nx1 vector. The grant for next year is given by

$$B_{t+1} = X\hat{\varphi}_t,$$

where B_{t+1} is an mx1 vector. Thus, the grant equals the predicted welfare expenses $\hat{Y}_t$ according to the regression model (7). Here, we ignore the effect of Z on the size of the macro budget.[20] Note that the grant received by local government i at time t+1 thus depends not only on X_i, but on both spending need determinants X and inefficiency levels Z at time t of *all* local governments.

In order to determine the effect of the allocation model on the equilibrium inefficiency level Z_t^* we need to analyze the relevant FOC. This is similar to the FOC (4) we derived for the problem P_t in Section 4.2. In matrix notation, with some abuse of notation, we now have

---

[18] We assume for simplicity that the regression model includes the correct set of exogenous variables X_j, j=1,...,n.

[19] Orthogonal is not the same as uncorrelated. Orthogonal means that the scalar product (or inner product) is 0; uncorrelated means that the scalar product of the vectors' centered (mean corrected) forms is 0. This is the same if both vectors are centered and have mean zero, which is not the case here.

[20] We ignore the fact that the predicted expenses may not sum to exactly the same amount as the actual expenses. Including this would imply scaling, i.e. multiplying each element of the vector B_{t+1} by the same number, which is determined exogenous of the model.

(8) $$L^d(Z_t) - C^d(Z_t \mid X) = \gamma\iota - \delta B^d_{t+1}(Z_t).$$

Here, L^d, C^d and B^d are mx1 vectors. The superscript d denotes the derivative with respect to the variable between brackets. E.g., the i-th element of L^d is the derivative of L with respect to Z, evaluated in $Z_{i,t}$, and the i-th element of $B^d_{t+1}(Z_t)$ is the derivative of $B_{i,t+1}$ with respect to $Z_{i,t}$. The variable ι is an mx1 vector of ones.

Thus, we need to take the derivative of the grant $B_{i,t+1}$ with respect to $Z_{i,t}$ (for each local government i). In the expression for $B_{i,t+1}$, $Z_{i,t}$ enters only via $\hat{\varphi}_t$, and in the expression for $\hat{\varphi}_t$ itself $Z_{i,t}$ enters only via $Y_{i,t}$. In order to obtain the derivative of $B_{i,t+1}$ with respect to $Z_{i,t}$, note that the derivative of the vector Y_t with respect to $Z_{i,t}$ is a vector which has zeros everywhere except for the i-th element, which equals γ. Thus, the derivative of $\hat{\varphi}_t$ with respect to $Z_{i,t}$ equals γ times the i-th column of the matrix $(X'X)^{-1}X'$, and the derivative of $B_{i,t+1}$ with respect to $Z_{i,t}$ equals the i-th row of X multiplied by this γ times the i-th column of the matrix $(X'X)^{-1}X'$. Thus, we have the following.

Proposition 6: With the regression method we have $B^d_{i,t+1}(Z_{i,t}) = \gamma\left[X(X'X)^{-1}X'\right]_{ii} \equiv \gamma h_{ii}$ and the block grant system increases efficiency if and only if $\delta h_{ii} < 1 - \alpha$.

How can this result be interpreted? First note that the effect of increased efficiency of local government i on its next period's grant depends not only on its own exogenous spending need determinants X_i, but also on those of the other local governments. However, the derivative does *not* depend on (any) inefficiency levels.

We can also see from Proposition 6 that local government i's grant $B_{i,t+1}$ reacts strongly to its inefficiency level $Z_{i,t}$ if (and only if) h_{ii}, the i-th diagonal element of $X(X'X)^{-1}X'$, is large (in absolute value). The matrix $X(X'X)^{-1}X'$ is known as the projection matrix or hat matrix. It transforms observed values Y_t into predicted values

$\hat{Y}_t$, given the regression equation (7): $\hat{Y}_t = X\hat{\varphi}_t = X(X'X)^{-1}X'Y_t$. The diagonal elements of this hat matrix, h_{ii}, can be interpreted as leverages. They describe the influence of an observation on the predicted value for that observation. A high value of h_{ii} means that the observation $Y_{i,t}$ is influential in determining $\hat{Y}_{i,t}$. It is well known (e.g., Hoaglin and Welsch, 1978) that $0 \leq h_{ii} \leq 1$, and that the average value equals n/m, where n is the number of parameters (here: exogenous spending need determinants) and m the number of observations (here: local jurisdictions). Clearly, if $\hat{Y}_{i,t}$ is determined to a relatively large extent by $Y_{i,t}$, then the grant $B_{i,t+1}$ is determined to a relatively large extent by $Z_{i,t}$. Again, the inequality in the proposition compares the present value of the eventual block grant increase resulting from spending an additional euro on welfare, δh_{ii}, to the grant increase due to spending one more euro when a matching grant is in place $(1-\alpha)$.

For municipality i, the FOC (8) now becomes

$$(9) \qquad L^d(Z_{i,t}) - C^d(Z_{i,t} \mid X) = \gamma(1-\delta h_{ii}).$$

As we assume that parameters γ and δ do not differ between local governments, the RHS of the FOC is constant. Local government i increases efficiency (i.e., $Z_{i,t}^* < Z_{i,0}^*$) after the introduction of the new system if and only if $\delta h_{ii} < 1-\alpha$. This is particularly likely to be the case if α and δ are small, and the observation $Y_{i,t}$ is not too influential (h_{ii} is small, which is generally the case if $m >> n$). Recall that with a fixed block grant, the RHS of the FOC, representing the marginal costs from welfare spending, would equal γ. The regression-based allocation system yields lower marginal costs because of adjustments deemed necessary out of equity concerns.

We now turn to the effects of the omitted variable problem. The effect of inefficiency on next period's grant, $B_{t+1}^d(Z_t)$, turns out to be independent of t (see Proposition 6). The specification of our model results in immediate transition to the new equilibrium efficiency level. In the real world, transition will not be immediate. Nevertheless, to be truly equitable it is desirable that the regression model (7) will converge to the true model (6) as local governments start working more efficiently as a result of the

incentives inherent in the block grant system. Thus, for the regression method to work well, the estimated parameter $\hat{\varphi}_t$ should converge to the true parameter β, and the grant B should converge to the 'fair' grant $X\beta$, at least in expected value. The expected value of the grant according to our model equals

$$\begin{aligned} EB_{t+1} &= EX\hat{\varphi}_t \\ &= E\left[X(X'X)^{-1}X'Y_t\right] \\ &= E\left[X(X'X)^{-1}X'\left(X\beta + Z_t\gamma + \mu_t\right)\right] \\ &= X\beta + X(X'X)^{-1}X'Z_t\gamma. \end{aligned}$$

Thus, the expected value of the grant at time t+1 equals the fair grant $X\beta$ plus an additional term, $X(X'X)^{-1}X'Z_t\gamma$, which depends on both spending need determinants and efficiency levels in all jurisdictions.

Proposition 7: Due to the omitted variable problem, under the regression method the estimated model does not converge to the true model, and the expected grant does not converge to the fair grant.

It is well known that the omitted variable problem affects the expected value of the estimated parameter ($\hat{\varphi}_t$), but not its variance. The omitted variable bias is given by[21] $E\hat{\varphi}_t - \beta = (X'X)^{-1}X'Z_t\gamma$. We have multiple regressors in X, and even if one of those is uncorrelated with Z_t, its estimate will be biased unless the regressor is uncorrelated with all other regressors too. In the current setting it seems reasonable to assume that the regressors are all correlated, so all estimates (all elements of the vector $\hat{\varphi}_t$) are biased. The bias is nonzero except in the special case where $Z_t = 0$, or where $X'Z_t$ is a vector of zero's, i.e., X and Z are orthogonal. The first case, $Z_t = 0$, implies complete efficiency in all jurisdictions, which is highly unlikely. The second case is highly unlikely mathematically, as in our model Z_t is determined by the FOC (9) as a function of X. Thus, convergence of the grant B to the fair grant $X\beta$ is highly

[21] Note that this bias equals γ times the slope from regressing Z_t on X.

unlikely. It is difficult to sign the omitted variable bias. Since all regressors in X can be pairwise correlated, it is next to impossible to obtain the direction of the biases.

5.4. Grant comparison

The first order condition describing the local government's efficiency choice sets the marginal benefit of welfare spending, $L'(Z_{i,t}) - C'(Z_{i,t} \mid X)$, equal to the marginal cost, i.e., net-of-grant welfare expenditures. Table 1 summarizes marginal costs for different grants, as derived above. They are constant for all grant types we study: they do not depend on the inefficiency level $Z_{i,t}$, but for some grant types they are different for municipalities with different values of budget share θ_i or leverage h_{ii}. In a Figure similar to Figure 2, the various marginal cost levels shown in Table 1 would be represented by horizontal lines. The inefficiency level $Z_{i,t}$ a municipality chooses decreases with increasing marginal cost of welfare spending, because higher marginal costs increase the incentive to work efficiently.

Whether a block grant gives municipalities a bigger incentive to work efficiently than a matching grant depends on parameter values. However, given that $\theta_i \in [0,1]$ and $h_{ii} \in [0,1]$, Table 1 shows that a previous period shares block grant gives a smaller efficiency incentive than block grant where shares are fixed or regression-based, except in extreme cases.

Table 1. Marginal cost of welfare spending under different grants (RHS of FOC)

Grant type	**Marginal cost of welfare spending**
Fixed block grant (or no grant)	γ
Matching grant	$\alpha\gamma$
Fixed shares block grant	$(1 - \delta\theta_i)\gamma$
Previous period shares block grant	$(1 - \delta)\gamma$
Regression-based block grant	$(1 - \delta h_{ii})\gamma$

Table 2 summarizes the efficiency effects of replacing a matching grant with a block grant. It shows that the fixed shares block grant is more efficient than the matching grant if $\delta < \frac{1-\propto}{\theta_i}$. The intuition behind this is straightforward. Efficiency improves if a local government's increase in welfare expenditures by one euro results in a smaller

grant increase under the block grant system than under the matching grant system. With a matching grant, spending an additional euro results in 1 – α of extra grant money. Spending one additional euro under a fixed shares block grant results in θ_i euro extra next year. The present value of that is $\delta\theta_i$. So, the fixed shares block grant is more efficient than the matching grant if $\delta\theta_i < 1-\propto$, or $\delta < \frac{1-\propto}{\theta_i}$. Note that the denominator reflects the increase in next year's grant resulting from spending more under a fixed shares block grant. With a previous period shares block grant, spending one additional euro results in one euro in extra grant money next year. Now, the denominator becomes one, and the previous period shares block grant is more efficient than the matching grant if $\delta < 1-\propto$. With regression-based shares, spending an additional euro results in h_{ii} euro in extra grant money next year, and the regression-based block grant is more efficient than the matching grant if $\delta < \frac{1-\propto}{h_{ii}}$. Table 2 also includes two numerical examples to be discussed in the next subsection.

Table 2. Efficiency effects of replacing a matching grant with a block grant.

Block grant type	**Efficiency improves if and only if**	**α = 0.10**	**α = 0.25**
Fixed shares	$\delta < \frac{1-\propto}{\theta_i}$ (Proposition 4)	$\delta < 397$ (for average θ_i) $\delta < 7.5$ (for the highest θ_i)	$\delta < 331$ (for average θ_i) $\delta < 6.25$ (for the highest θ_i)
Previous period shares	$\delta < 1-\propto$ (Proposition 5)	$\delta < 0.9$	$\delta < 0.75$
Regression-based	$\delta < \frac{1-\propto}{h_{ii}}$ (Proposition 6)	$\delta < \frac{0.9}{h_{ii}}$ (average h_{ii}: $\delta < 13$)	$\delta < \frac{0.75}{h_{ii}}$ (average h_{ii}: $\delta < 11$)

In general, replacing a matching grant with a block grant improves efficiency if α and δ are sufficiently small. For a fixed shares block grant, efficiency additionally requires small θ_i, and for a regression-based grant, efficiency additionally requires small h_{ii}. The latter implies m >> n, or many jurisdictions and relatively few exogenous welfare spending need determinants. Policymakers can increase block grant efficiency by increasing the time lag between local governments' spending behavior and the resulting effect on grant size, e.g. using data for year *t–k* instead of *t*-1 for the

regression analysis, where k is an integer denoting the lag, $k>1$. Effectively, this replaces δ with $\delta^k<\delta$.

5.5. Application to the Netherlands

We now apply our results to the Dutch case. In the Netherlands, local governments originally received an open-ended matching grant to finance welfare spending, as described by the benchmark model above, with $\alpha = 0.25$.[22] In 2004, this was replaced by a system of block grants.

First, we compare the matching grant with a block grant where the shares θ_i in the macro budget for different local jurisdictions are fixed. In the Netherlands, the average value of θ_i equals 0.002.[23] According to Table 2, introducing a block grant with fixed shares increases efficiency in a municipality with an average value of θ_i if $\delta < 331$, which is easily satisfied – recall that in our model, $\delta \in [0,1]$. Still, the incentive to increase efficiency could be small in very big municipalities. The maximum value of θ_i is 0.12 (for Amsterdam). Thus, an efficiency increase in all, including the biggest, municipalities implies $\delta < 6.25$. Assuming $\delta = 0.95$, introducing a fixed shares block grant increases efficiency in every municipality if $\theta_i < 0.79 \,\forall i$, which will normally be the case. With high values of α, however, it is conceivable that introducing a block grant actually decreases efficiency in some large municipalities. For the Netherlands, this would require $\alpha \geq 0.89$ (again using $\theta_i = 0.12$ for Amsterdam), which is much higher than it has ever been. Thus, we can conclude that replacing the matching grant that existed in the Netherlands with a block grant with fixed shares would have increased efficiency in all municipalities. Such a block grant was not introduced, however.

Now consider previous period shares. Since 2004, the grant share of small municipalities in the Netherlands (< 25,000 inhabitants, where 9 percent of welfare recipients reside), depends on their share in welfare expenditures at t-2. The grant we

[22] In 2001 - 2003. Until 2001, $\alpha = 0.10$. We will not discuss this case in the text; results are given in Table 2.

[23] Calculated as 1/441, where 441 is the number of municipalities.

analyze in section 5.2 is based on the expenditure share in *t*-1 rather than *t*-2. Therefore, the actual effect of the Dutch grant system for small municipalities does not follow directly from Table 2. Also, the previous period shares system does not apply to large municipalities. If we nevertheless apply our results to the Dutch case, we find the following. Municipality *i* increases efficiency if a matching grant is replaced by a grant based on previous expenditure shares if and only if $\delta < 1-\propto$. With α equaling 0.25 before the matching grant was replaced, this requires a discount factor $\delta < 0.75$, which seems implausibly low. Even considering that the Dutch grant is actually based on the expenditure share in t-2 rather than t-1, so that we should use δ^2 instead of δ, this requires a low discount factor ($\delta < 0.87$). Thus, the new grant may have *reduced* efficiency for small municipalities. However, as large municipalities do have an incentive to reduce welfare dependency (see next paragraph), there is a downward pressure on the macro budget. As a result, for small municipalities, spending one additional euro actually results in less than one euro in extra grant money two years later. This improves their efficiency incentive somewhat.

For large municipalities ($\geq$ 40,000 inhabitants, where about 80 percent of welfare recipients live), regression-based grant allocation applies.[24] Policymakers implicitly assume that the estimated model converges to the true model as local governments start working more efficiently. Proposition 7 states that this is not to be expected. Although this method may well increase efficiency relative to the old matching grant system, full efficiency is unlikely to obtain and the expected grant will remain biased. For a Dutch municipality with an average value of h_{ii} (which is 0.07),[25] the regression-based grant is more efficient than the matching grant if $\delta < \frac{1-\propto}{0.07}$ (Table 2). With $\alpha = 0.25$ this implies $\delta < 11$. As $\delta \in [0,1]$ the average municipality has increased efficiency. However, some municipalities may have disproportionate influence on

[24] The grant of medium sized municipalities is determined partly by their share in the previous period, and partly by regression results. The importance of both components depends on the number of inhabitants: with increasing size, regression results increase in importance.

[25] In 2009. Calculated as *n*, the number of exogenous spending need determinants (14), divided by *m*, the number of large and medium sized municipalities (205).

their estimated welfare expenditures. With a reasonably safe value of 0.95 for δ,[26] efficiency requires $h_{ii} < 0.79$, which only doesn't hold for extreme outliers (recall $0 \leq h_{ii} \leq 1$).[27] For the extreme case where $h_{ii} = 1$, the regression-based grant is equal to the previous period shares block grant, and the efficiency condition is the same as well (see Table 2).

We can conclude that, according to our model, replacing the matching grant with a regression-based block grant in the Netherlands has increased efficiency in all municipalities concerned. This is in line with empirical evidence. Preliminary estimates of the effect of the introduction of block grants on the number of welfare recipients point to a reduction between 8 (Van Es and Van Vuuren, 2010) and 15 percent (Kok et al., 2007). However, the introduction of the previous period block grant for small municipalities may have reduced efficiency there.

Our modeling framework allows for a theoretical analysis of another aspect of the Dutch system. In the Netherlands, a municipality's grant is not allowed to deviate too much from actual welfare expenditures. We extend our model in this direction in Appendix B. Overall, the results for the equilibrium inefficiency level are ambiguous, but numerical simulations indicate that the *ex ante* limit may well increase inefficiency by limiting the marginal cost of spending.

6. Conclusion

This paper discusses the use of regression-based block grants for a welfare system with decentralized administration but centralized financing. With uniform benefits and eligibility rules, welfare grants from the central government to local jurisdictions should be designed in such a way that they provide local governments with the right incentives to work efficiently, that is, give benefits only to those who really need it.

---

[26] Note that as it takes time for data to become available and for regression analyses to be carried out, the time lag in the Netherlands is usually bigger than one year (2-3 years). As a result, we are actually assuming here that δ^2 or δ^3 is 0.95, which is rather on the safe side.

[27] As a rule of thumb in regression analysis, values exceeding two or three times the average value of h_{ii} (here: 0.14 or 0.21) are considered influential outliers that merit close inspection, and, possibly, exclusion from the analysis (e.g., Hoaglin and Welsch, 1978).

Block grants are preferred over matching grants because of their efficiency, but they usually have the disadvantage of being less equitable. Such grants may be insufficiently low for local jurisdictions with high exogenous welfare spending needs. In this paper we consider a block grant allocation system which tries to avoid this disadvantage by letting grants depend on expected spending need. Such grants were introduced in the Netherlands in 2004.

Dutch policymakers use econometric techniques to forecast future spending needs from a regression of welfare expenditures on observable exogenous spending need determinants. With grant shares derived from such a regression, a block grant should ensure that local governments that operate reasonably efficiently will not need to use own resources to finance welfare expenditures. Because total grants add up to no more than forecasted aggregated welfare expenditures, excess spending is discouraged. In this way, the Dutch aim to enjoy the benefits of block grants (efficiency), without the disadvantage associated with them (inequity). However, since inefficiency is not observed, the regression has an omitted variable problem and thereby a bias. We derive the size of this bias. In contrast to what policymakers claim, we show that in our simplified setting the regression model does not converge to the true model and the grant does not converge to the fair grant due to the omitted variable bias.

A second problem with the regression method is that it gives rise to perverse incentives. Matching grants reduce the marginal cost of welfare spending and thereby increase the attractiveness of working inefficiently. Standard block grants do not have this property. However, the regression-based block grants discussed here have the property that higher expenditures increase future grants. This provides perverse incentives to local administrators by lowering the marginal net-of-grant costs of welfare spending. We show that full efficiency is not likely to obtain with a regression-based block grant. In extreme cases, efficiency may be even lower than under a matching grant system for relatively large local governments, for which expenditures usually have a greater effect on future grants than for small ones. So, in general, this type of block grant will result both in less than perfect efficiency, and in continuing disparities in efficiency across local governments of different size.

We analyze the efficiency results of replacing a matching grant with a regression-based block grant, and with two alternative block grants. We conclude from our model that the introduction of regression-based grants in *large* Dutch municipalities has improved efficiency there. The reason is that with many local jurisdictions relative to the number of exogenous spending need determinants, the perverse incentive turns out to be small. However, our analysis suggests that the introduction of a block grant with shares depending on previous period shares in *small* Dutch municipalities may have decreased efficiency, whereas a block grant with fixed shares would have had the opposite effect.

We conclude that the Dutch style regression-based block grant may be successfully applied by countries wishing to combine local administration, central financing, and efficient administration of welfare, while ensuring uniform eligibility and benefit levels and an equitable welfare burden for local jurisdictions. The method may also be applied to other programs. Our analysis shows under which conditions such regression-based grants may improve efficiency.

Acknowledgements

We thank Wouter Vermeulen, Stephen Ferris, participants of the 2011 annual meeting of the Public Choice Society (San Antonio, Texas) and of ASSET 2010 (Alicante), and seminar participants at the University of Groningen for useful comments.

References

Akai, N., Sato, M., 2008. Too big or too small? A synthetic view of the commitment problem of interregional transfers, Journal of Urban Economics, 64, 551-559.

Akai, N., Silva, E.C.D., 2009. Interregional redistribution as a cure to the soft budget syndrome in federations, International Tax and Public Finance 16: 43-58.

Allers, M.A., 2012. Yardstick competition, fiscal disparities, and equalization, Economics Letters 117: 4-6.

Baicker, K., 2005. Extensive or intensive generosity? The price and income effects of federal grants, The Review of Economics and Statistics 87, 371-384.

Bird, R. M., Smart, M., 2002. Intergovernmental Fiscal Transfers: International Lessons for Developing Countries, World Development 30, 899–912.

Blank, R.M., 2002. Evaluating welfare reform in the United States, Journal of Economic Literature 40, 1105-1166.

Boadway, R., 2006. Intergovernmental redistributive transfers: efficiency and equity, in: Ahmad, E., Brosio, G. (Eds.), Handbook of Fiscal Federalism, Edward Elgar , Cheltenham.

Boadway, R., Flatters, F., 1982. Efficiency and equalization payments in a federal system of government: a synthesis and extension of resent results, Canadian Journal of Economics 15, 614-633.

Bradbury, K., Zhao, B., 2009. Measuring non-school fiscal disparities among municipalities, National Tax Journal LXII, 25-56.

Bramley, G., 1990. Equalization Grants and Local Expenditure Needs. The Price of Equality, Avebury, Aldershot.

Breuillé, M.-L., Gary-Bobo, R.J., 2007. Sharing budgetary austerity under free mobility and asymmetric information: an optimal regulation approach to fiscal federalism. Journal of Public Economics 91, 1177–1196.

Breuillé, M.-L., Madiès, T., Taugourdeau, E., 2010. Gross versus net equalization scheme in a federation with decentralized leadership, Journal of Urban Economics 68, 205-214.

Buchanan, J.M., 1950. Federalism and fiscal equity, The American Economic Review 40, 583-599.

Buchanan, J.M., 1952. Federal grants and resource allocation, The Journal of Political Economy 60, 208-217.

Buchanan, J. and Goetz, C., 1972. Efficiency limits of fiscal mobility: An assessment of the Tiebout model, Journal of Public Economics 1, 25–43.

Bucovetsky, S., 1998. Federalism, equalization and risk aversion, Journal of Public Economics 67, 301–328.

Chernick, H., 1998. Fiscal effects of block grants for the needy: An interpretation of the evidence, International Tax and Public Finance 5, 205-233.

Cornes, R.C., Silva, E.C.D., 2002. Local public goods, interregional transfers and private information. European Economic Review 46, 329–356.

Duncan, A., Smith, P., 1996. Modelling local government budgetary choices under expenditure limitation, Fiscal Studies 16, 95-110.

Gilbert, G., Rocaboy, Y., 1996. Local public spending in France: The case of welfare programmes at the département level, in: Pola, G., France, G., Levaggi, R., Developments in Local Government Finance, Edward Elgar, Cheltenham.

Goodspeed, T.J., 2002. Bailouts in a federation, International Tax and Public Finance 9, 409-421.

Hayek, F. A., 1945. The Use of Knowledge in Society, American Economic Review 35: 519-30.

Hoaglin, D.C., Welsch, E.E., 1978. The hat matrix in regression and ANOVA, The American Statistician 32, 17-22.

Huber, B., Runkel, M., 2006. Optimal design of intergovernmental grants under asymmetric information. International Tax and Public Finance 13, 25–41.

Kok, L., Groot, I., Güler, D., 2007. Kwantitatief effect WWB, SEO, Amsterdam.

Köthenbürger, M., 2004. Tax competition in a fiscal union with decentralized leadership, Journal of Urban Economics 55, 498-513.

Ladd, H.F., 1994. Measuring disparities in the fiscal condition of local governments, in: J.E. Anderson (Eds.), Fiscal equalization for state and local government finance, Greenwood, 21-54.

Le Grand, J., 1975. Fiscal equity and central government grants to local authorities, The Economic Journal 85, 531-547.

Oates, W. E., 1972. Fiscal Federalism. Harcourt Brace Jovanovich, New York.

Oates, W. E., 1999. An Essay on Fiscal Federalism, Journal of Economic Literature 37, 1120–1149.

Raff, H., Wilson, J.D., 1997. Income redistribution with well-informed local governments. International Tax and Public Finance 4, 407–427.

Ribar, D.C., Wilhelm, M.O., 1999. The demand for welfare generosity, The Review of Economics and Statistics 81, 96-108.

Riordan, M.H., Sappington, D.E.M., 1988. Optimal contracts with public ex post information, Journal of Economic Theory 45, 189–199.

Shleifer, A., 1985. A theory of yardstick competition, Rand Journal of Economics 16, 319-327.

Strausz, R., 2006. Interim Information in Long-Term Contracts, Journal of Economics & Management Strategy, 15, 1041–1067.

Van Es, F., Van Vuuren, D.J., 2010. Minder uitkeringen door decentralisering bijstand, Economisch Statistische Berichten 95 (4589), July 9.

Von Hagen, J., 2006. Achieving economic stabilization by sharing risk within countries, in: Boadway, R., Shah, A. (Eds.), Intergovernmental fiscal transfers. Principles and practice, The World Bank , Washington, DC.

Wildasin, D., 1986. Urban Public Finance. Harwood, New York.

Appendix A: A penalty on deficits

Contrary to our assumption so far, it is conceivable that the local government's objective function puts a greater weight on a welfare deficit ($Y_t > B_t$) than on a surplus. Van Es and Van Vuuren (2010) provide evidence that such an asymmetry prevails for welfare financing in Dutch municipalities. The reason could be that the municipality has a decision maker – say, the alderman for social services – who is in charge of welfare administration, and who maximizes his own objective function. This objective function takes into account that as long as there is some excess budget, the decision maker can (to some extent) go his own sweet way. However, if there is a deficit, the other aldermen will be involved. The deficit may, e.g., lead to a – politically painful – tax increase. A large deficit may even force a local government to reduce welfare spending, because such a deficit is difficult to finance. Our basic model does not account for this. There, welfare spending depends on the marginal costs and benefits, which are independent of the deficit or surplus on welfare finance. That may not be realistic. Indeed, in the Dutch case, if large municipalities reduce welfare spending because of the incentives inherent in the regression-based grant allocation, the grants to small municipalities will decrease because of a lower macro budget. This may force them to reduce welfare spending too.

Consider an individual municipality and ignore the subscript i for simplicity. The objective function for problem P_t, equation (2), now becomes

$$\sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[\Lambda\{B_\tau(Z_{\tau-1}) - Y_\tau - C(Z_\tau \mid X)\} + L(Z_\tau) \right]$$

where $\Lambda(x)$ is an increasing, continuously differentiable function with $\Lambda(0)=0$ and which is steeper for negative values of x than for positive values. For example, we could let $\Lambda(x)$ be defined as follows: $\Lambda(x) = 1/4 - (x - 1/2)^2$ for $x < 0$ and $\Lambda(x) = x$ for $x \geq 0$. Now, monetary costs and benefits are weighed more heavily in case of a deficit, while the weight of the non-monetary benefit L remains the same.

The FOC (4) now becomes

$$L'(Z_t) - C'(Z_t \mid X)\Lambda'_t(\cdot) = \gamma\Lambda'_t(\cdot) - \delta B_{t+1}'(Z_t)\Lambda'_{t+1}(\cdot),$$

where $\Lambda'_t(\cdot)$ denotes the derivative of the function Λ evaluated in $B_t(Z_{t-1}) - Y_t - C(Z_t \mid X)$. This equation is difficult to solve for Z_t. However, we can observe the following. The term $\Lambda'_t(\cdot)$ depends on Z_t but also on Z_{t-1}. This implies that we can no longer solve for Z_t independent of time. Instead, there will be some sort of adjustment process.

Suppose that for $x \geq 0$ we have $\Lambda(x) = x$, so $\Lambda'(x) = 1$, as before. In this case, it can easily be seen from the FOC that the penalty on deficits has two effects. First, if there is a deficit at time t, money in this period has a greater weight in the objective function. This, however, applies both to the cost term C and to the expenditure Y. The overall effect on Z_t is ambiguous and depends on the relative size of C' and γ, as the FOC shows. If spending an additional euro on administrative costs (fraud prevention, active labor market policy) results in a welfare spending decrease bigger than one euro, getting in deficit will give an incentive to improve efficiency. In this case, a lower macro budget because of efficiency improvements elsewhere could give a municipality an (extra) incentive to work more efficiently.

Second, if there is going to be a deficit at time t+1, that period's budget has a greater weight in the objective function. The municipality will respond by choosing higher inefficiency at time t in order to avoid a low budget in the next period. In fact, even in case of excess budget the decision maker has less of an incentive to increase efficiency, because this reduces his future grant and thereby may increase the likelihood of future deficits.

Appendix B: Upper bounds on deficits and surpluses

As mentioned in Section 3, in order to insulate local governments from budgetary shock too great to cope with, in the Netherlands the difference between the block grant and actual welfare expenditures is limited both *ex post* and *ex ante*. The *ex post* limit fixes the upper limit of the (positive) difference between actual welfare expenditures and the block grant allocated to a municipality in the same year. If, at the end of the year, welfare expenditures turn out to exceed the grant by an amount of more than 10 percent of expenditures, the municipality receives additional funding *ex post* which finances the additional deficit. This *ex post* deficit limit affects roughly two dozen (out of 441) municipalities.[28] In practice, the *ex ante* limit is more important. It has been binding for more than half of all municipalities every year since the introduction of the new grant system. Differences (in absolute value) between the grant allocated to a municipality and welfare expenditures in year *t*-2 are subject to an upper limit *ex ante* of 7.5 percent of welfare expenditures. As of 2009, structural deficits (at least 2.5 percent during three consecutive years) are subject to an upper limit *ex post* too.

Ex post deficit limit

The *ex post* deficit limit fixed the upper limit of the (positive) difference between actual welfare expenditures and the block grant allocated to a municipality in the same year. Now, the grant B_t satisfies the condition

$$\frac{Y_t - B_t}{B_t} \leq \omega^p.$$

[28] In 2007 – 2011, this number increased considerably; in 2009, 171 municipalities were affected. In this period, the macro budget was not adapted annually as described in section 3. Instead, the municipalities had agreed to bear the risk of increasing welfare expenditures themselves. With hindsight, this rather unfortunately coincided with a severe economic downturn. As from 2012, the mechanism described in section 3 is in place again.

Here, $\omega^p \in (0,1)$ (recall that in the Netherlands, $\omega^p = 0.1$). This means that if, at the end of the year, welfare expenditures turn out to exceed the grant by an amount of more than $100\omega^p$ percent of the grant, the municipality can apply for additional funding *ex post* which finances the deficit in excess of $\omega^p B_t$. However, this additional funding is granted only if the deficit cannot be blamed on local administrators.[29]

With this restriction, the objective function for problem P_t, (2), becomes

$$\sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[\max\left\{ B_\tau(Z_{\tau-1}) - Y_\tau, -\omega^p B_\tau(Z_{\tau-1}) \right\} - C(Z_\tau \mid X) + L(Z_\tau) \right].$$

This implies that the FOC (4) may now change: the RHS may change during some periods, depending on whether or not the restriction is binding at time t or t+1 (or both). Thus, we can distinguish two possible effects of the restriction. Clearly, at any point in time, one, both or neither of these effects may occur.

First, if the restriction is binding (only) at time t, the term γ drops out of the FOC (4). Equilibrium in Figure 2 is at the intersection with the curve $-\delta B_{t+1}'(Z_t)$, which is below $\gamma - \delta B_{t+1}'(Z_t)$ as γ>0. This indicates that the municipality will now choose higher Z_t. The intuition for this effect is as follows. Higher inefficiency means higher welfare expenditures, but now at some point this implies no additional *own* expenditures, as the match rate jumps from zero to one as soon as the restriction kicks in.

The *ex post* limit does not apply if the deficit is a result of the local government's own policy. However, a municipality may expect a weak local labor market which is likely to lead to a deficit exceeding $100\omega^p$ percent of the grant in the following year. If (and only if) the restriction is binding at time t+1, the term $-\delta B_{t+1}'(Z_t)$ vanishes from the FOC (4). Equilibrium in Figure 2 is at point E_2 instead of E_1. Intuitively, higher

[29] In the Netherlands, an independent commission investigates whether a municipality's own policies have caused the deficit. In that case, it must bear the deficit itself.

efficiency (lower Z_t) normally results in a lower grant for the next period, but if the grant is supplemented to cover additional welfare expenditures anyway, this negative effect of greater efficiency drops out. In this case, lowering Z_t does not reduce next year's grant anymore. Thus, the municipality would choose lower Z_t, but only to the extent that the restriction is still binding at time t+1.

So, on the one hand, the upper bound for the deficit may limit the cost of working inefficiently (in terms of welfare expenditures paid out of own resources). On the other hand, however, it may take away part of the problem that municipalities are reluctant to increase efficiency because this lowers their grant in the next period.

Ex ante limit

In the Netherlands, the *ex ante* limit applies to medium-sized and large municipalities, for which the grant B_t is determined (at least in part) by the regression method. The idea behind it is that the difference between the grant B_t and a (hypothetical) grant based on expenses in the previous year (equal to $Y_{t\text{-}1}$, see Section 5.2) should not be too large. Thus, the *ex ante* limit effectively fixes the upper limit of the difference (in absolute value) between the grant B_t allocated to a municipality and welfare expenditures in year t-1, $Y_{t\text{-}1}$, *ex ante* at $100\omega^a$ percent of the grant, with $\omega^a \in (0,1)$:[30]

$$\left| \frac{Y_{t-1} - B_t}{B_t} \right| \leq \omega^a .$$

Again, this condition puts an additional restriction on the variables for the maximization problem (2). However, the effect on the equilibrium efficiency level Z_t^* is more difficult to assess than in the case of the *ex post* limit.

The objective function for problem P_t (equation (2)) can now be written as

[30] In the Netherlands, $\omega^a = 0.075$. To be precise, the Dutch system compares the grant to expenditures in the year t-2 (rather than t-1). We use t-1 here for expositional convenience.

$$\sum_{\tau=t}^{\infty}\delta^{\tau-t}\left[\max\left\{\min\left\{B_\tau(Z_{\tau-1}),\frac{1}{1-\omega^a}Y_{\tau-1}\right\},\frac{1}{1+\omega^a}Y_{\tau-1}\right\}-Y_\tau-C(Z_\tau\mid X)+L(Z_\tau)\right].$$

If the *ex ante* limit is binding, the FOC (4) is affected. The *ex ante* limit may be binding as a result of an exogenous shock, but may also become binding as a result of the municipality's own policy (choice of Z_t). This seriously complicates the analysis and prevents us from deriving general results on the effects of the *ex ante* limit on efficiency. For example, the inefficiency level Z_t that a municipality would want to choose without the *ex ante* limit might be such that the *ex ante* limit applies. But then the municipality might want to adjust Z_t to avoid this (for example, the municipality may want to avoid that its grant B_{t+1} is subject to a maximum). Or alternatively, a municipality might want to adjust its inefficiency level Z_{t-1} to make the *ex ante* limit apply at time t+1 (for example, to force a lower bound on its grant). This yields many possibilities and it is virtually impossible to compare all of those. In the remainder of this appendix we therefore focus on how the constraint affects the FOC and thereby efficiency, taking as given that it is binding. The idea behind this approach is that in this way at least we can obtain some impression of the possible effect on efficiency of this *ex ante* limit, assuming that it will be binding sometimes. Recall that in the Dutch case, the *ex ante* limit has been binding for more than half of all municipalities every year since the introduction of the block grant system.

The FOC (4) is affected only if the constraint is binding at time t. In that case, the term $-\delta B_{t+1}'(Z_t)$ is replaced by a term $-\frac{\delta\gamma}{1\pm\omega^a}$, where the '+' applies if the grant hits the lower bound (deficit) and the '-' if it hits the upper bound (surplus):

$$L'(Z_t)-C'(Z_t\mid X)=\gamma\left(1-\frac{\delta}{1\pm\omega^a}\right).$$

In these cases, the marginal cost of spending (the RHS of the FOC) thus becomes a constant.

In order to be able to draw conclusions on the direction of the change in the marginal cost, and thereby on the effect of the *ex ante* limit on efficiency, we need to specify

the function $B_{t+1}(Z_t)$. Consider plausible values of δ and ω^a: δ = 0.95, ω^a = 0.075 (the value used in the Netherlands). It is easy to verify that when the *ex ante* limit kicks in, in case of a deficit the marginal cost (RHS) equals approximately 0.12γ, lower than under the matching grant with α = 0.25 (in which case the RHS equals $\alpha\gamma$ = 0.25γ). With the regression-based block grant, which is used for the Dutch municipalities where the *ex ante* limit applies, the RHS equals $(1 - \delta h_{ii})\gamma = 0.93\gamma$ for average leverage h_{ii}, which is even higher. As values exceeding two or three times the average value of h_{ii} (here: 0.14 or 0.21) are considered unusual (e.g., Hoaglin and Welsch, 1978), we may conclude that this will usually hold true. Thus, in case of a deficit the *ex ante* limit results in lower marginal cost of spending, and therefore in greater inefficiency. Note that for the given parameter values, in case of a surplus the *ex ante* limit implies an RHS that is slightly negative, which in our setting results in the maximum inefficiency level $Z^{\max}$. Overall, this indicates that the *ex ante* limit may well increase inefficiency by limiting the marginal cost of spending.

Figures

Figure 1. Municipality budget constraint before (AC) and after (DE) the reform.

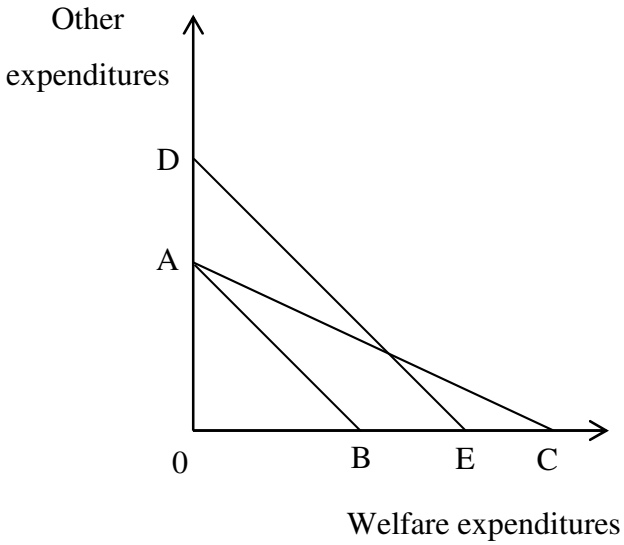
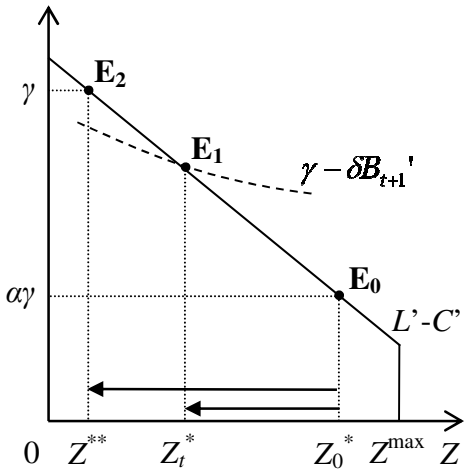


Figure 2. Solution of the model.



List of research reports

12001-HRM&OB: Veltrop, D.B., C.L.M. Hermes, T.J.B.M. Postma and J. de Haan, A Tale of Two Factions: Exploring the Relationship between Factional Faultlines and Conflict Management in Pension Fund Boards

12002-EEF: Angelini, V. and J.O. Mierau, Social and Economic Aspects of Childhood Health: Evidence from Western-Europe

12003-Other: Valkenhoef, G.H.M. van, T. Tervonen, E.O. de Brock and H. Hillege, Clinical trials information in drug development and regulation: existing systems and standards

12004-EEF: Toolsema, L.A. and M.A. Allers, Welfare financing: Grant allocation and efficiency

www.rug.nl/feb